Evaluating the Performance of Large Language Models on GAOKAO Benchmark

Year: 2023 | Citations: 151 | Authors: Xiaotian Zhang, Chun-yan Li, Yi Zong

---

Abstract

Large Language Models(LLMs) have demonstrated remarkable performance across various natural language processing tasks; however, how to comprehensively and accurately assess their performance becomes an urgent issue to be addressed. This paper introduces GAOKAO-Bench, an intuitive benchmark that employs questions from the Chinese GAOKAO examination as test samples, including both subjective and objective questions. To align with human examination methods, we design a method based on zero-shot settings to evaluate the performance of LLMs. With human evaluation, we obtain the converted total score of LLMs, including GPT-4, ChatGPT and ERNIE-Bot.Our findings reveal that LLMs have achieved competitive scores in Chinese GAOKAO examination, while they exhibit significant performance disparities across various subjects. We also use LLMs to grade the subjective questions, and find that model scores achieve a moderate level of consistency with human scores. In conclusion, this research contributes a robust evaluation benchmark for future large language models and offers valuable insights into the advantages and limitations of such models.